Gene-Environment Interaction Analysis using GPU

Daniel Berglund

December 2014

Outline

Problem and Aim

- Genetic and environment factors are know to affect the risks of diseases
- Interaction can exist between these factors
- Logistic Regression can be used to search for interaction
  - Iterative method
  - Models hidden probabilities of the outcomes as a linear model

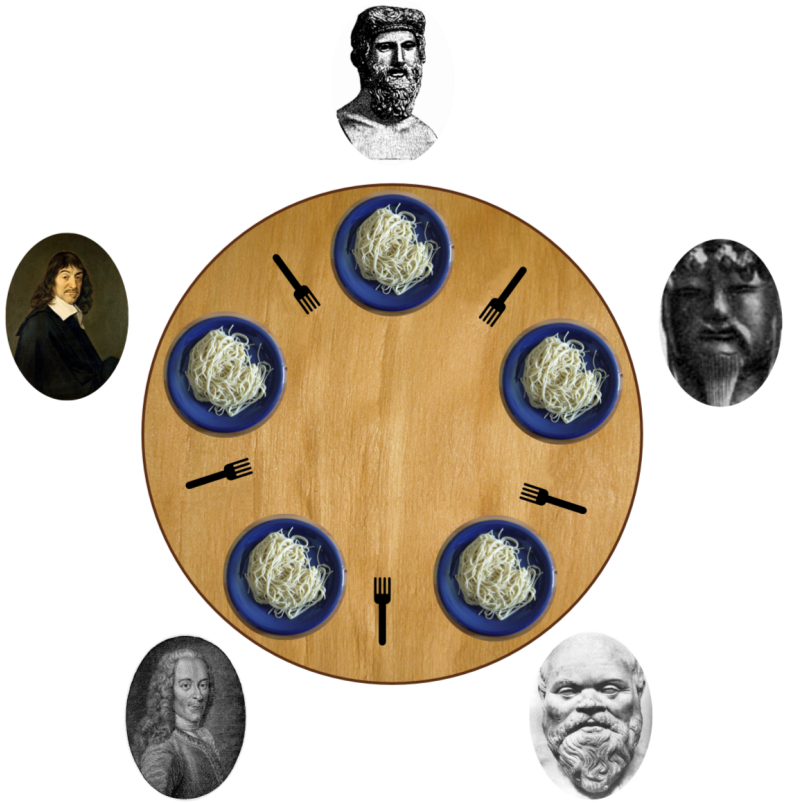
Problem and Aim

- Data amounts are increasing
- Need for more speed
- GPUs have previously shown good results for gene-gene interaction

Recoding

- The measures for additive interaction are defined for positive odds ratios
- Can be adjusted by recoding
- Recoding switches the reference group with the group with lowest risks
- Guarantees that $OR \geq 1$

Concurrency and Dinning Philosophers



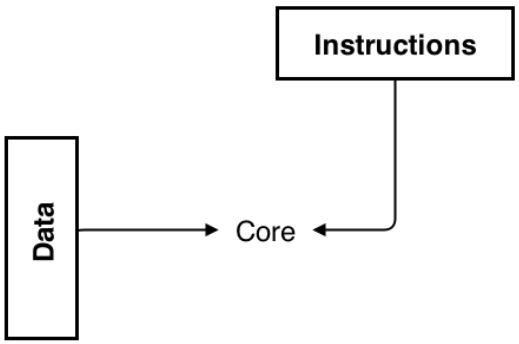
A Possible Solution

- Think
- Wait for a fork to become available and pick up that fork
- Wait for and pick up the other fork
- Eat
- Put down the forks one by one
- Go back to thinking

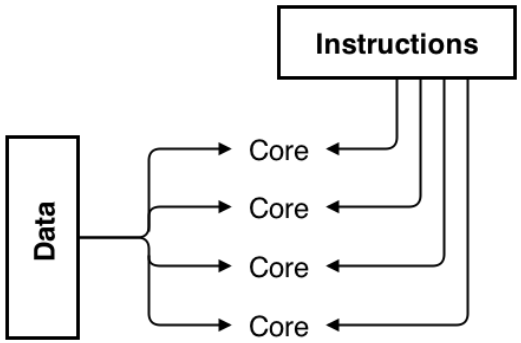
Deadlock

- Stuck when everyone holds one fork
- This is a deadlock
- Can be solved with a mediator

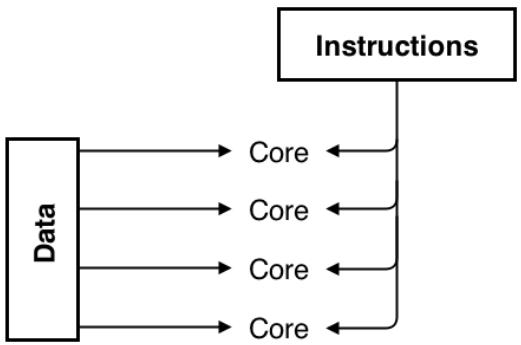
Concurrency Architectures



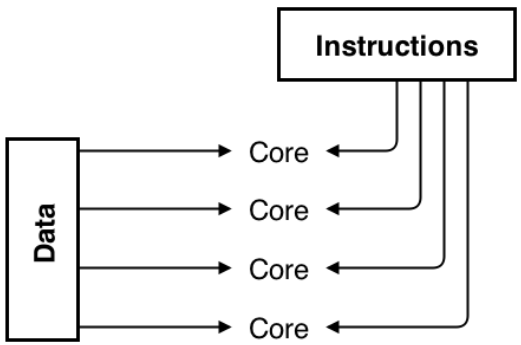
SISD



MISD

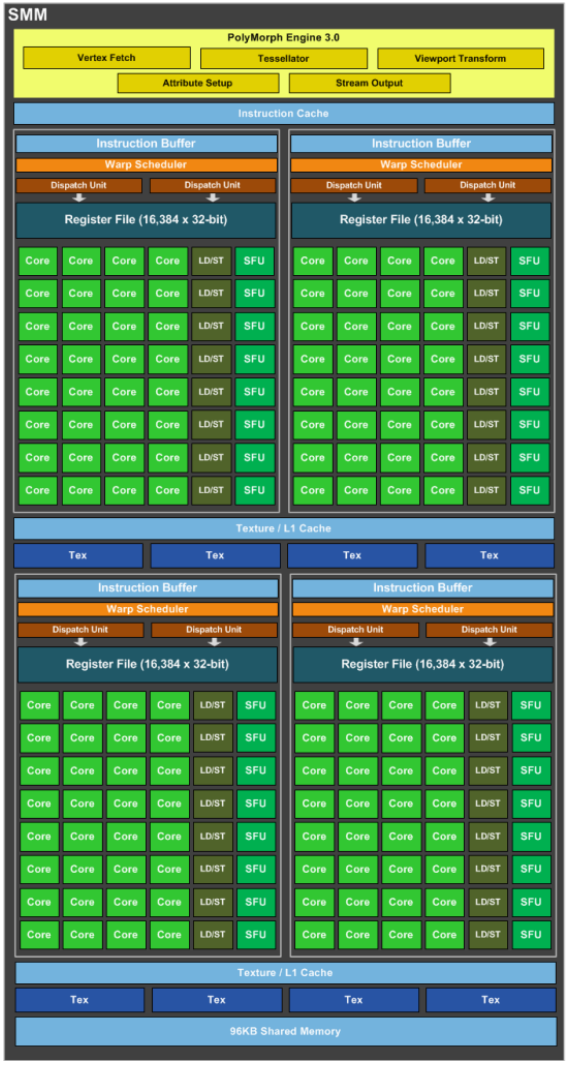


SIMD



MIMD

GPU Architecture



GPU Architecture

- Separate memory from the normal RAM
- Data has to be explicitly transferred to and from the GPU
- Simpler cores than CPU cores
- Single instruction, multiple thread(SIMT)

CUDA

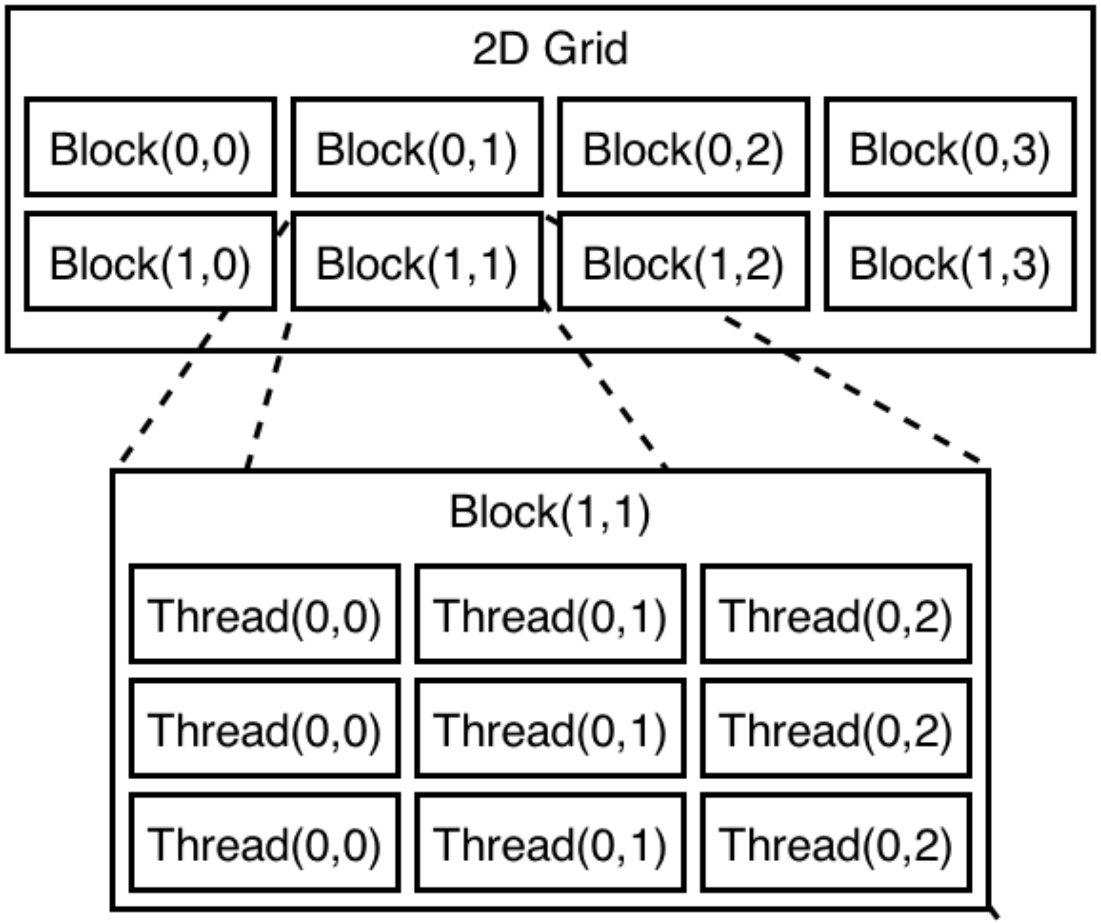
- For NVIDIA GPUs by NVIDIA
- Supported by various libraries, CUBLAS, ArrayFire, and so on
- Calculations done by kernels

Kernel Example

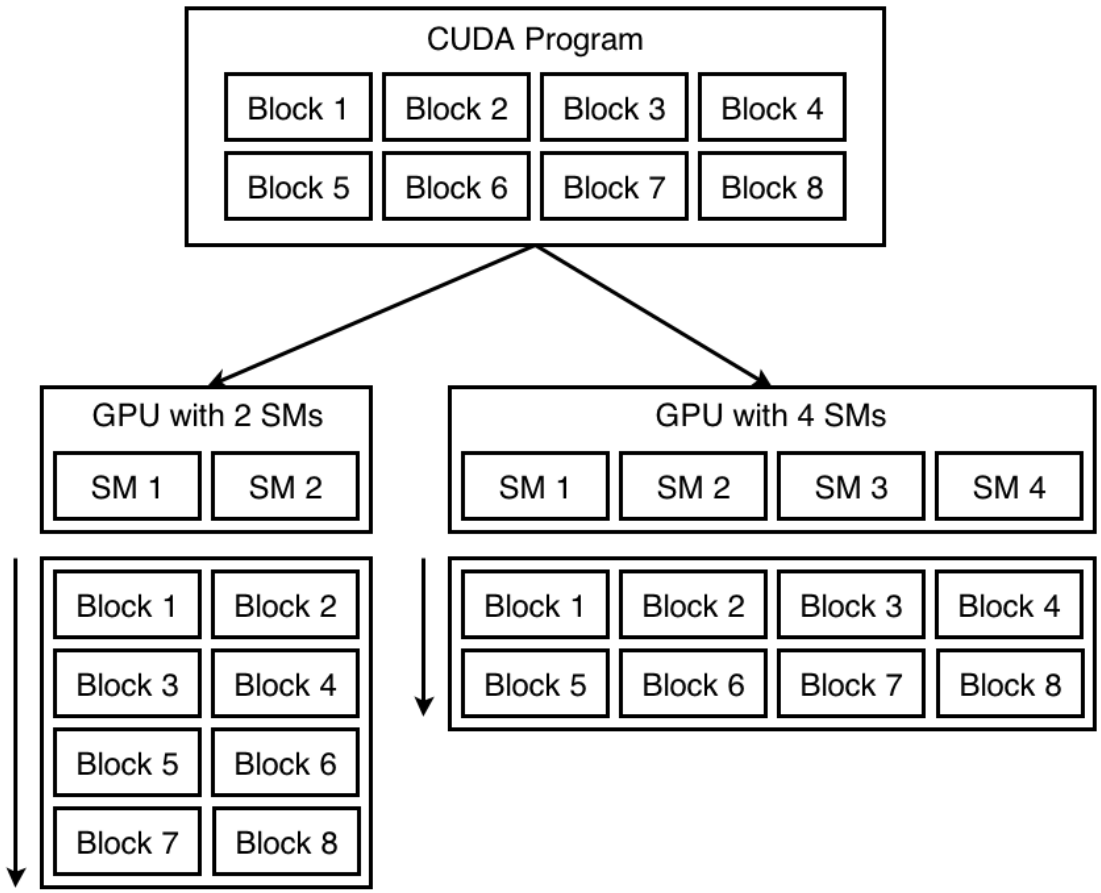
```
__global__ void Add(float* A, float* B, float* C)
{
    int i = threadIdx.x;
    C[i] = A[i] + B[i];
}

Add<<< N,M >>>(A, B, C);
```

Blocks


2D Grid
Block(0,0)
Block(0,1)
Block(0,2)
Block(0,3)
Block(1,0)
Block(1,1)
Block(1,2)
Block(1,3)
Block(1,1)
Thread(0,0)
Thread(0,1)
Thread(0,2)
Thread(0,0)
Thread(0,1)
Thread(0,2)
Thread(0,0)
Thread(0,1)
Thread(0,2)

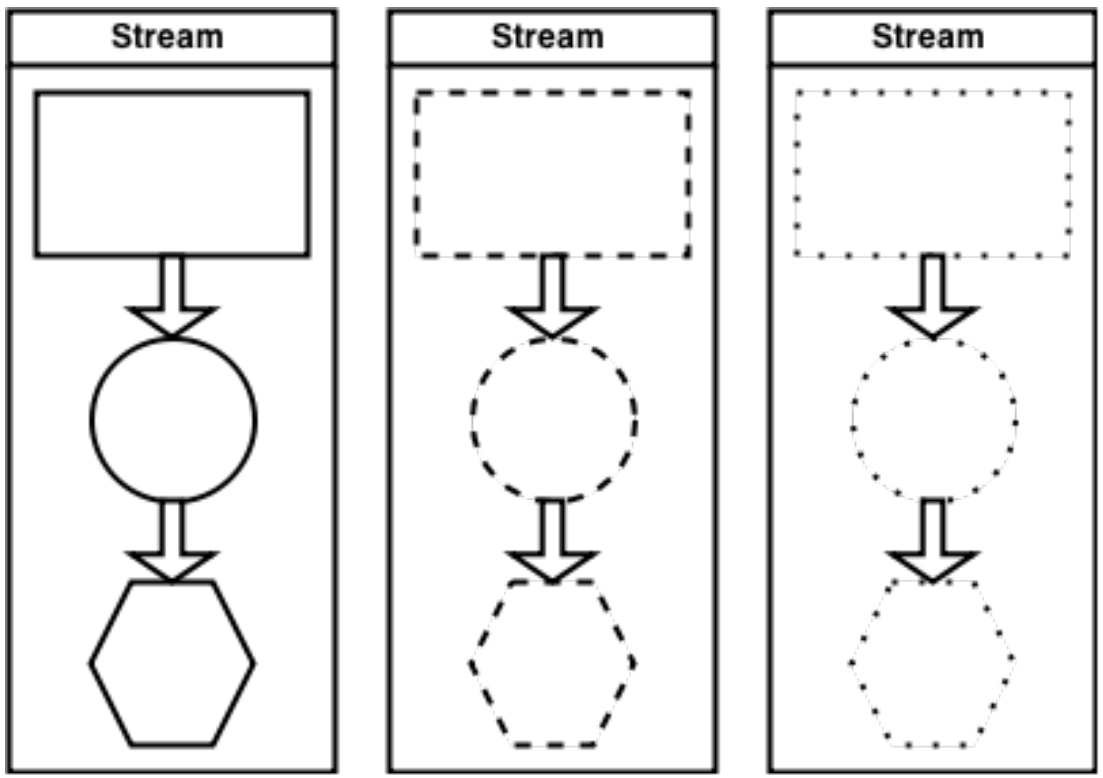
Blocks


CUDA Program
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
GPU with 2 SMs
SM 1
SM 2
GPU with 4 SMs
SM 1
SM 2
SM 3
SM 4
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8

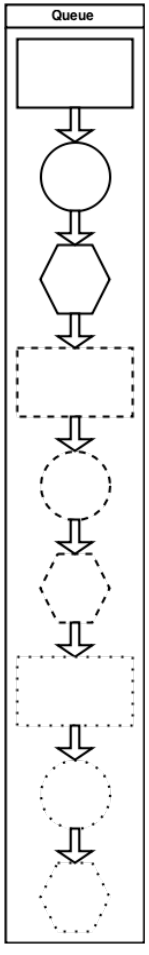
Streams

- Kernels and transfers can be performed on streams
- Transfers and kernels can overlap when using streams

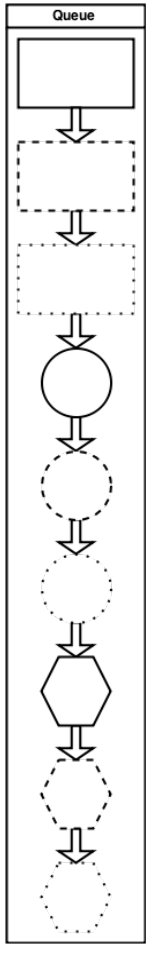
Streams



Streams



Looped over stream



Looped over kernel

Agile Development and Clean Code

- Code in small units, commonly classes
- Unit testing, test the units in isolation
- Mocks, i.e. fake objects
- If a unit is working incorrectly only its corresponding unit test should fail
- Integration tests can be used to test if the units work properly together

Wrappers

- Wrappers are used to “fix” interfaces
- Does not perform the task, delegates it

Wrappers

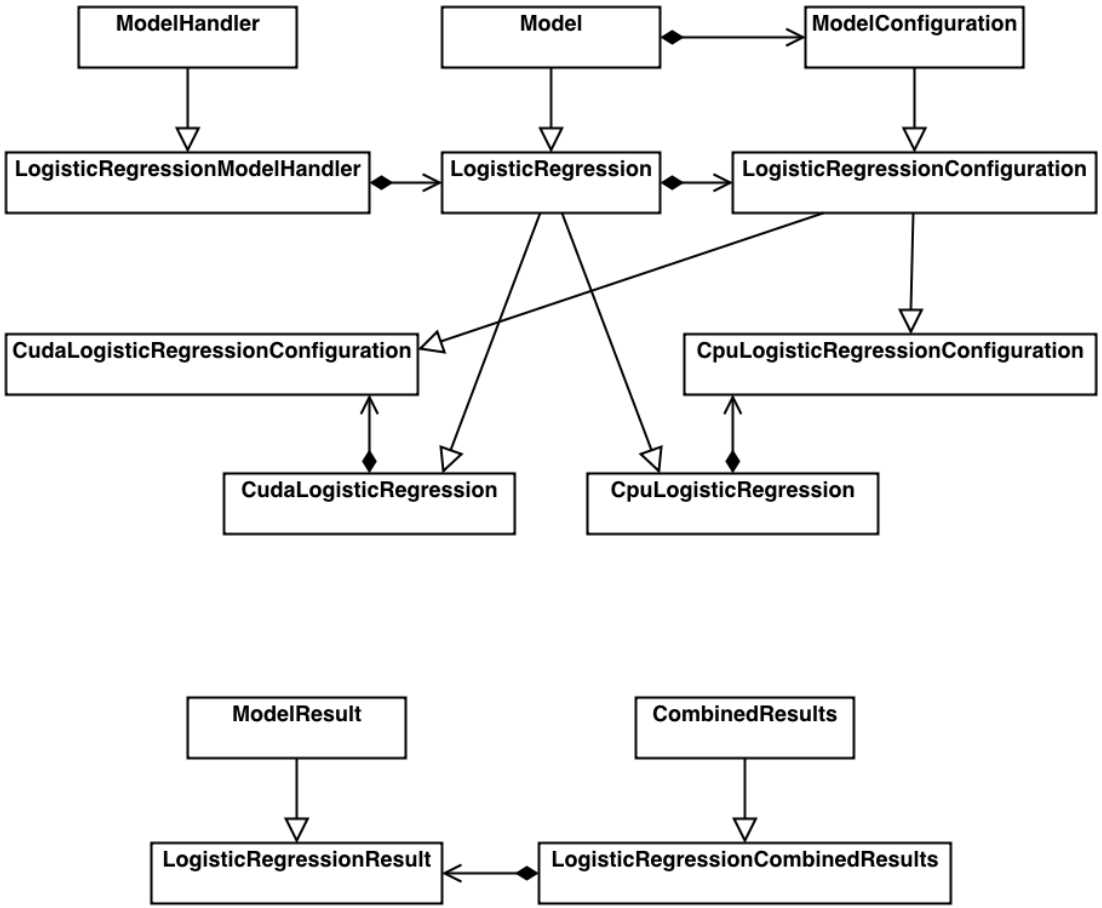
cublasSgemv(*cublasHandle*, *transpose*, *m*, *n*, α, *matrix*, *ld_matrix*, *vector*, *inc_vector*, β, *result*, *inc_result*);

matrixVectorMultiply(*matrix*, *vector*, *result*, α, β);

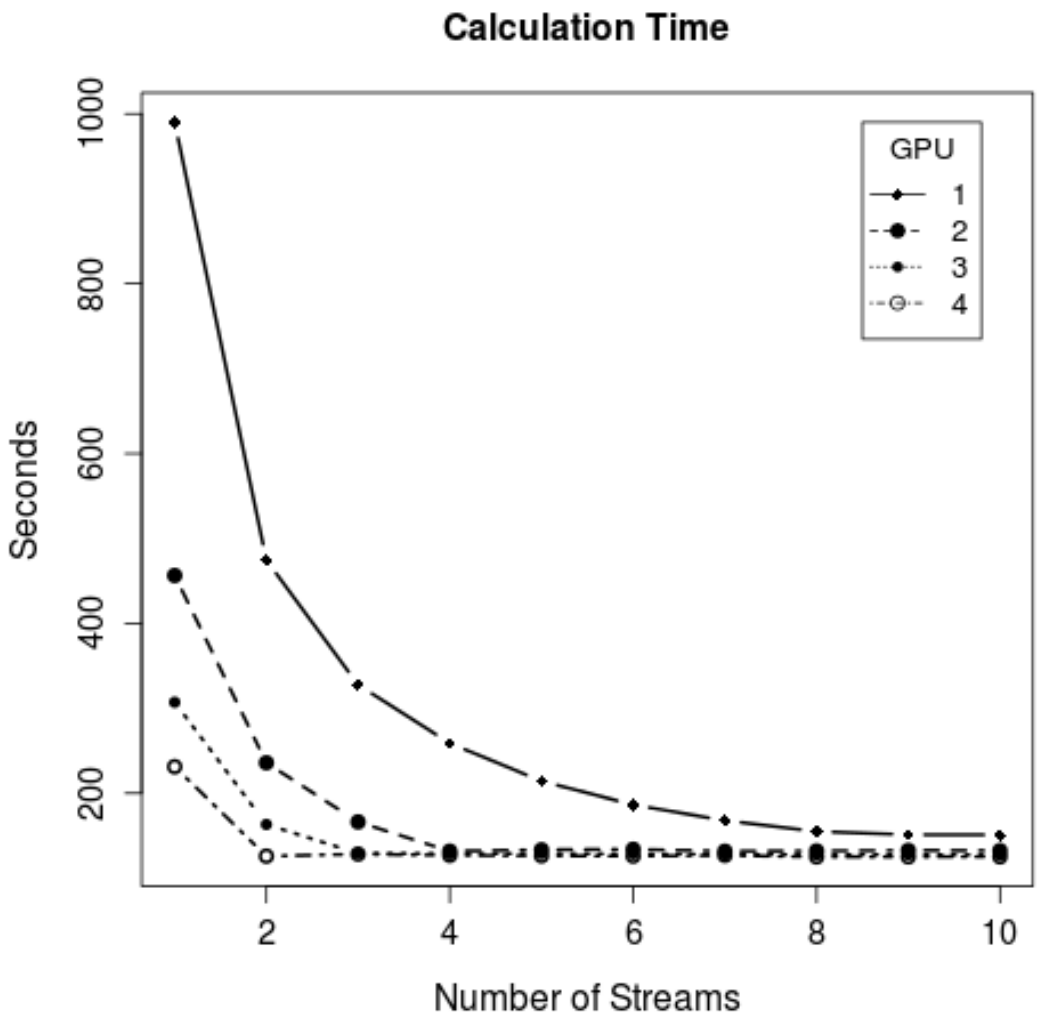
Structure of Program

- Modularised
- Almost full unit test coverage
- Handles each gene environment combination independently
- Threads fetches a genetic factor from a queue
- Each thread corresponds to a stream on a GPU
- Some parts are done using CPU

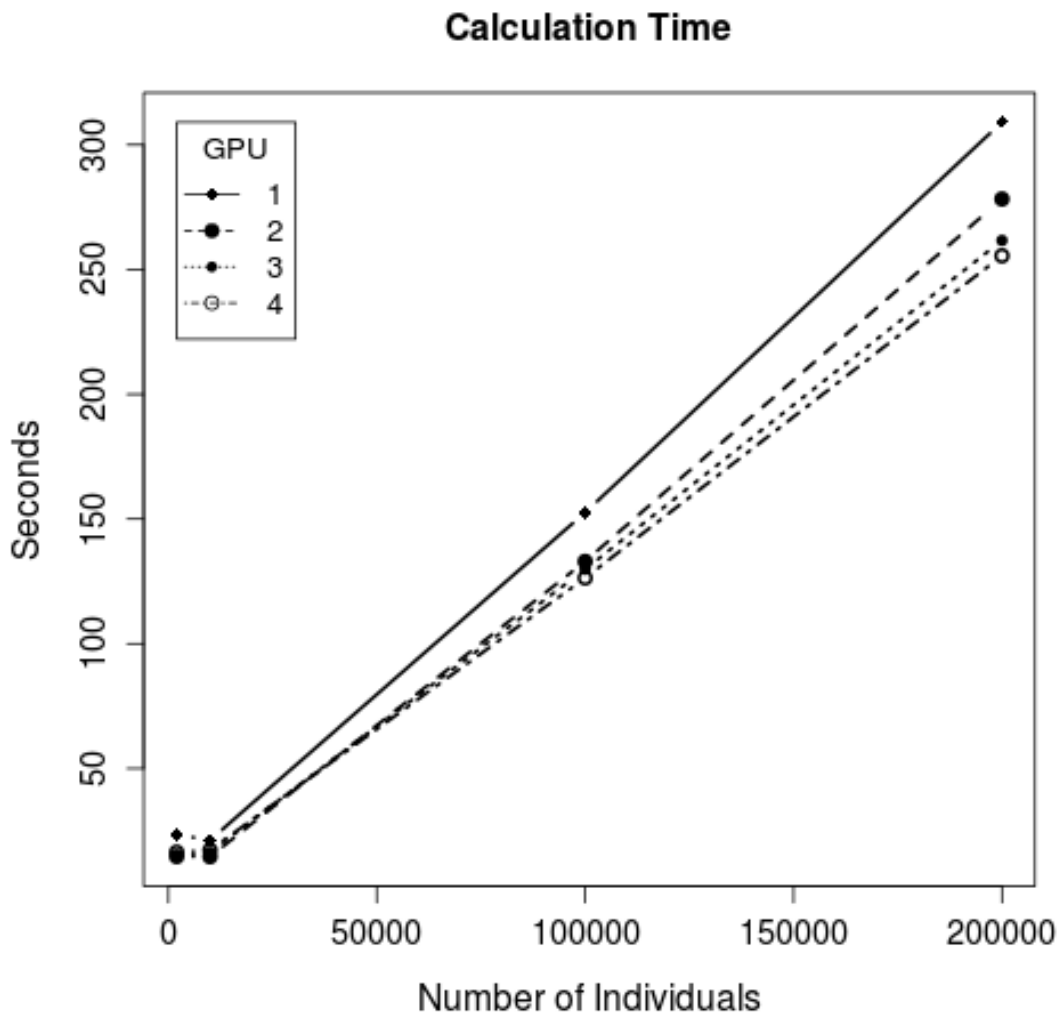
Model


ModelHandler
Model
ModelConfiguration
LogisticRegressionModelHandler
LogisticRegression
LogisticRegressionConfiguration
CudaLogisticRegressionConfiguration
CpuLogisticRegressionConfiguration
CudaLogisticRegression
CpuLogisticRegression
ModelResult
CombinedResults
LogisticRegressionResult
LogisticRegressionCombinedResults

Streams


Calculation Time
Seconds
1000
800
600
400
200
2
4
6
8
10
Number of Streams
GPU
1
2
3
4

Saturated Streams



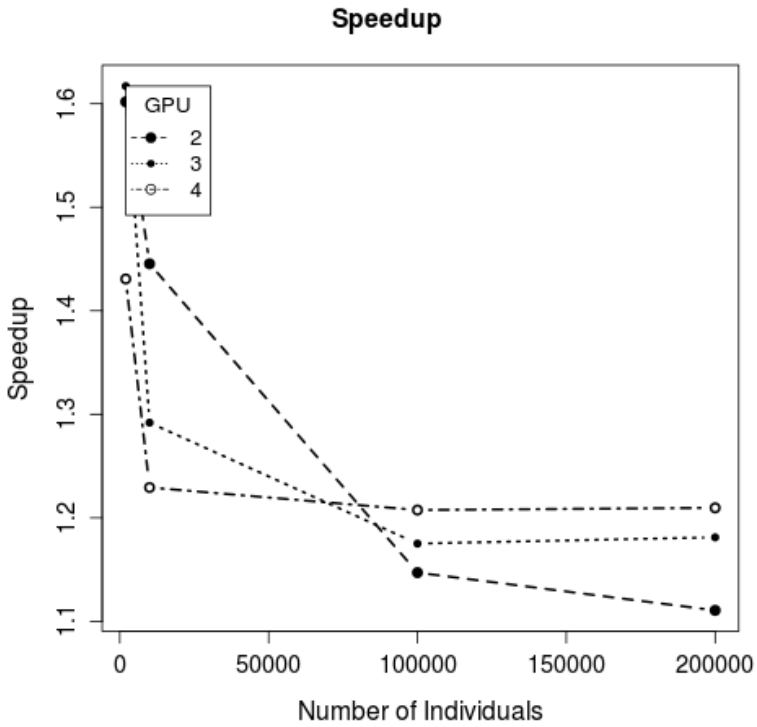
Speedup and Efficiency

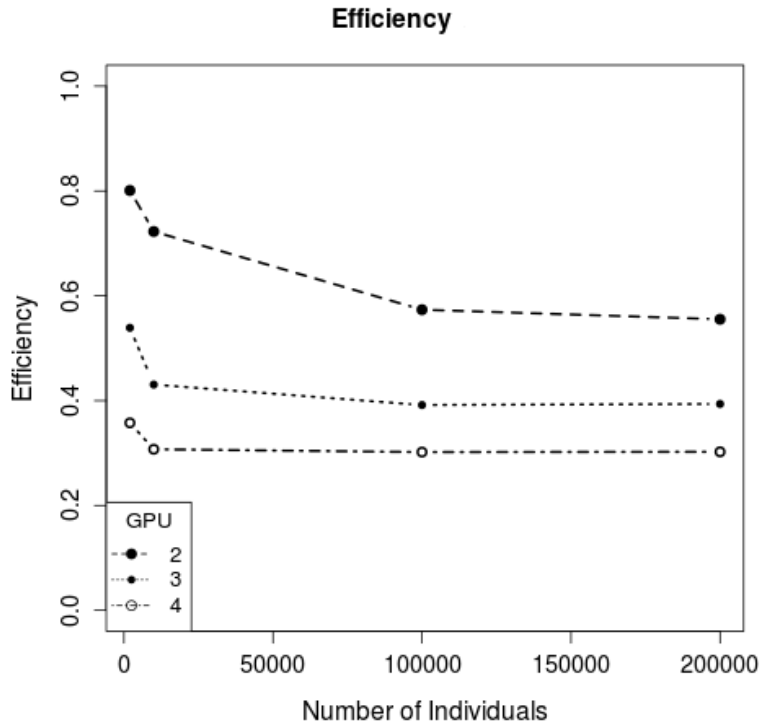
$$S(p) = \frac{T(1)}{T(p)}$$

$$E(p) = \frac{S(p)}{p} = \frac{T(1)}{pT(p)}$$

Speedup and Efficiency

Versus one GPU



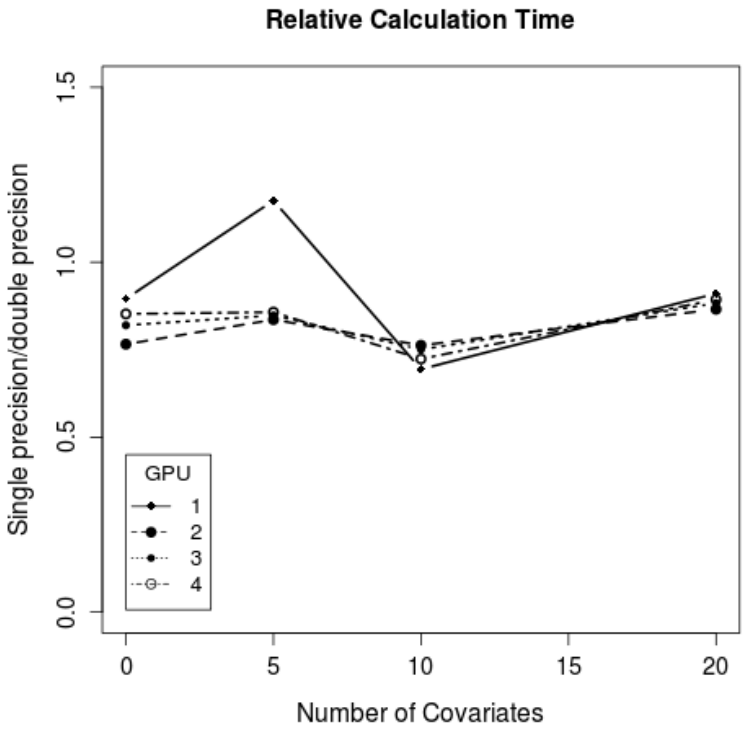


Time Distribution for Kernels

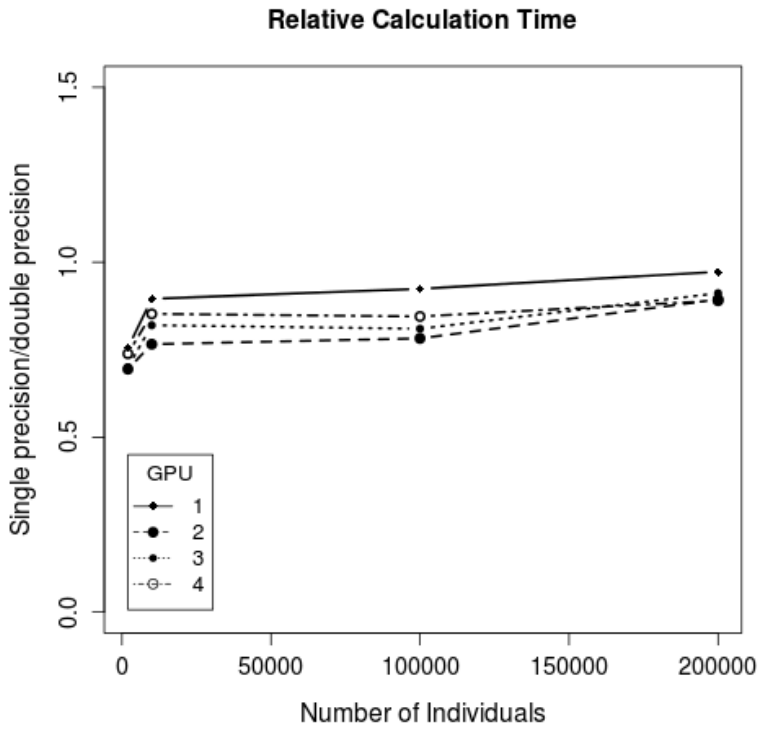
$*$ is element by element multiplication

Covariates Individuals		0 10k	0 100k	20 10k	20 100k
$M1^T \cdot V1$	CUBLAS	66.6	72.2	43	53.9
$M1 \cdot M2$	CUBLAS	18.9	14.4	36.1	25.8
$M1 \cdot V1$	CUBLAS	7.1	8.1	5.9	7.4
$V1 * V2$	Custom	2.9	2.3	11.8	10.6
$\frac{e^{V1}}{1+e^{V1}}$	Custom	1.1	0.9	0.7	0.7
$V1 * \log(V2) + (1 - V1) * \log(1 - V2)$	Custom	0.8	0.6	0.6	0.5
$V1 - V2$	Custom	0.7	0.6	0.5	0.4
$V1 * (1 - V1)$	Custom	0.7	0.5	0.4	0.4
Dot product	CUBLAS	0.7	0.3	0.3	0.2
$\Sigma V1$	CUBLAS	0.5	0.1	0.2	0.1

Single Versus Double Precision

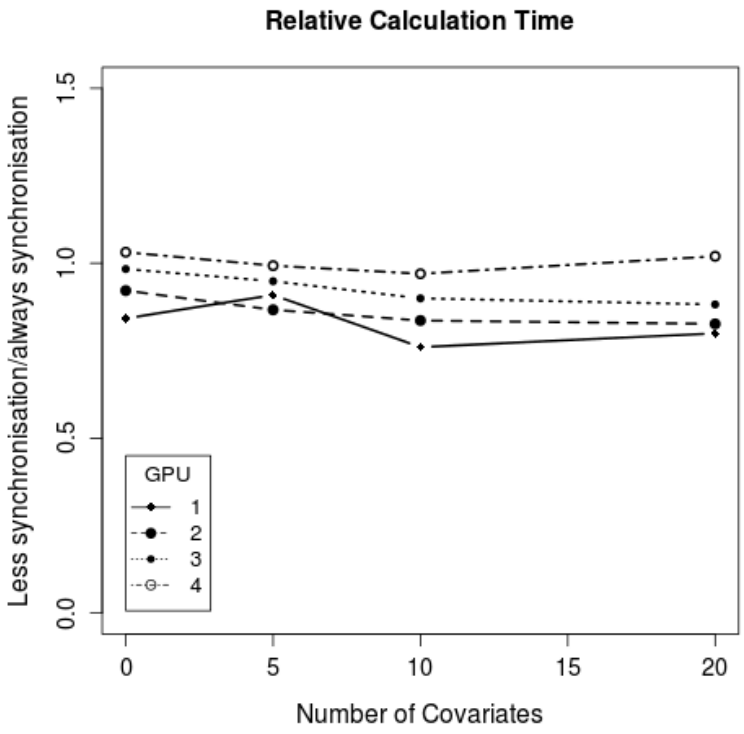


10000 individuals

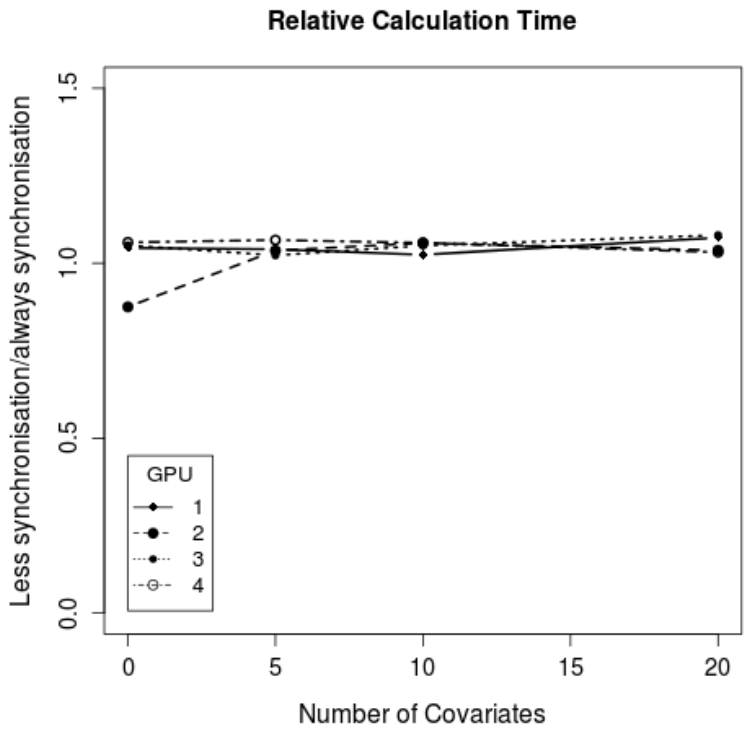


0 covariates

Syncing

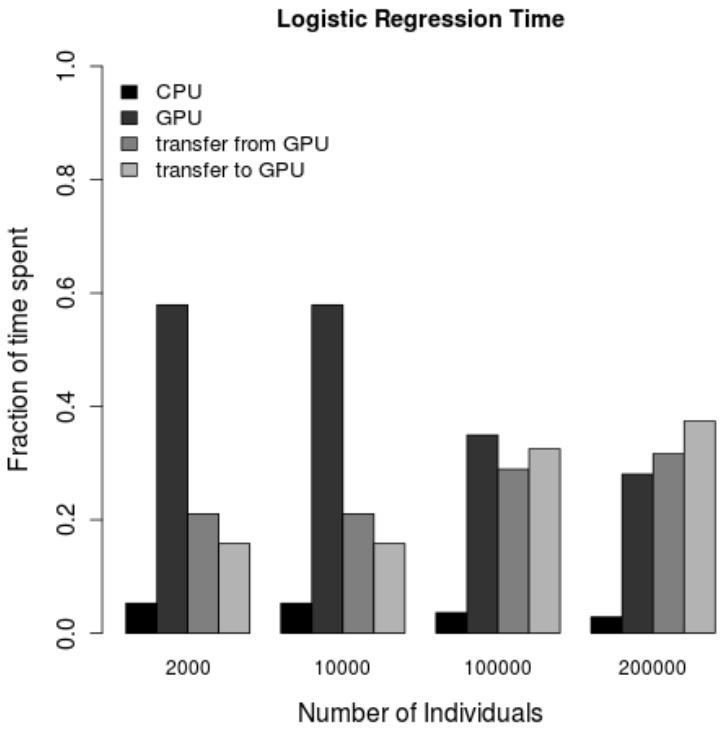


2000 individuals

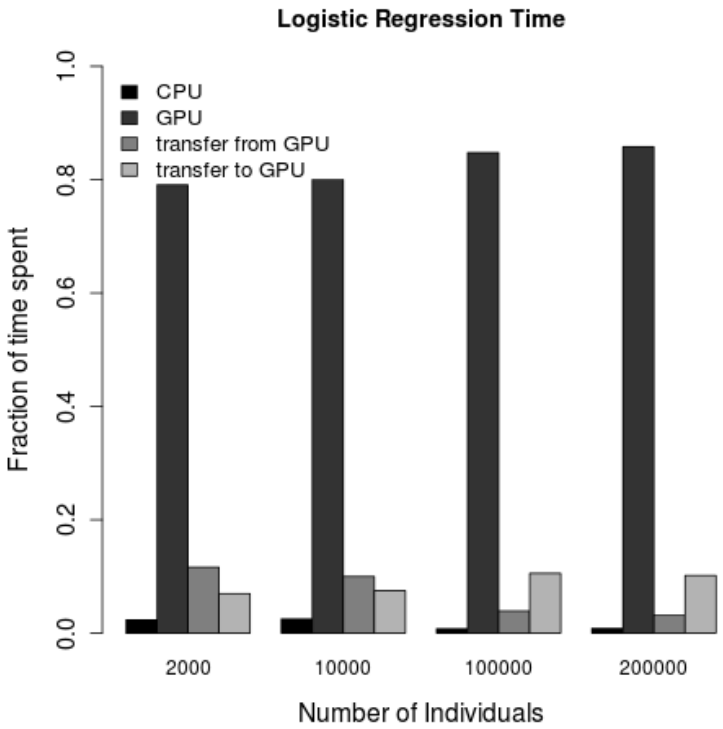


200 000 individuals

Synchronisation



Less synchronisation



Always synchronise

Conclusions

- GPUs suits well for interaction, depending on method
- Use one, perhaps two, GPUs with the program
- Single precision gives better performance than double precision
- Increased synchronisation increases performance slightly when number of individuals is above 100 000

Outlook

- Moving more parts to the GPU could fix the bad scaling
- Clusters for more speed
- Need for better statistics and definitions of interaction for non binary factors
- For binary factors look into gene-gene interaction methods if further speed is needed

Matrix Decomposition

- Pseudo inverse, A^+, is defined for general matrices
- Can be found by using singular value decomposition

$$A = U\Sigma V^T$$

$$A^+ = V\Sigma^+ U^T$$

Odds, Odds Ratio and Additive Interaction

$$\Omega = \frac{\pi}{1-\pi}$$

$$\theta = \frac{\Omega_1}{\Omega_2}$$

$$\theta_{\textit{both factors present}} > \theta_{\textit{first factor present}} + \theta_{\textit{second factor present}} - 1$$